辞書の見出し語集合と代表性

佐藤 理史
名古屋大学大学院工学研究科電子情報システム専攻
ssato@nuee.nagoya-u.ac.jp

1　はじめに

用語抽出研究や訳語推定研究の最終目標として、特定分野の用語集や対訳辞書の自動生成が掲げられることが多い。一見、これらの要素技術が十分な精度 (たとえば、99%以上) で実現できれば、用語集や対訳辞書の自動生成は達成できるように思える。もちろん、このような高い精度での要素技術の実現は容易ではなく、現実にはなかなか最終目標にたどり着かない。事実、これまでに自動的に生成されたもので、人間用の用語集や対訳辞書として実用的なレベルに達したものは皆無であると言ってよい。

しかしながら、どうやら別のところにも問題がありそうだ、というのが、本論文の出発点である。我々は、これまで、外国人名を対象とした対訳 (原綴とカタカナ訳) 辞書の自動編纂を目標に、要素技術の高度化を進めてきた [2, 3, 4, 5]。その結果として、大量の外国人名対訳データ (品質によってサイズが異なる；高品質データは約 35 万件) を保持するに至ったが、それらをそのまま束ねても、人間用の対訳辞書としては、おそらく機能しない。そのままでは、人間用の辞書として有効に機能するため「何か」が足りないのである。その「何か」とは何か、それを生み出すためにはどうしたらよいか。これらの問題に考察を加えることが、本論文の主題である。

2　辞書作成のモデル

辞書は、新聞、雑誌、書籍などに代表される情報パッケージの一つである。情報パッケージとは、「利用者とその目的を想定し、それに合致するように情報の収集、選択、加工を行なうこと」(編集＝情報のパッケージング) によって作り出された情報群の総体を意味する [6]。我々は、以前行なった「情報の自動編集」に関する研究において、このような考え方を提唱したが、その時点では、自動化できる部分、すなわち、素材情報の収集、選択、加工に力点を置いたため、それに先立つ設計は、提案した自動編集のモデルには含まれていない。しかしながら、本論文で扱う問題を議論するためには、この設計の部分を陽に記述する必要がある。

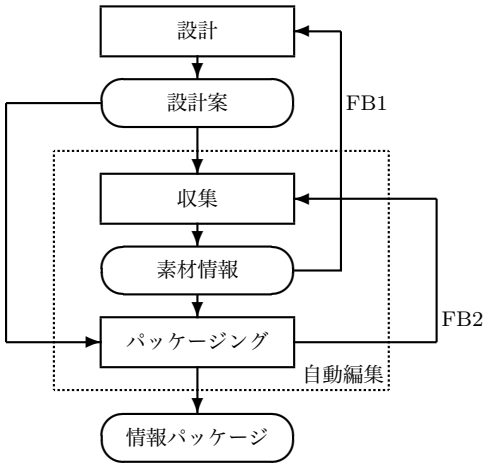

図 1: 情報パッケージの作成モデル

図 1 に、本論文で採用する「情報パッケージ作成」のモデルを示す。このモデルは、以前のモデルには存在しなかった、設計、その結果として得られる設計案、および、フィードバックループを構成するための 2 本の線 (FB1, FB2) が追加されている。

一般に、辞書は、見出し語とそれに対する内容記述から構成されるエントリーの集合としてモデル化できる。本論文では、内容記述の部分には立ち入らず、見出し語についてのみ議論する。辞書編纂の各段階において、見出し語に関わる処理は、次の通りである。

1. 設計：どのような語をどれだけ収録するかを定める
2. 収集：見出し語の候補となるものを集める
3. パッケージング：見出し語集合を具体的に定める (要素を列挙する)

3　見出し語集合と代表性

辞書において、その価値を規定する最も重要なものの一つは、見出し語集合である。辞書編纂者は、設計

の段階で「どのような語をどれだけ収録するか」を定める。しかし、利用者にとっては、実際に作成された見出し語集合が「集合としてどのような性質を持つか」ということの方が、より本質的となる。

辞書の利用者は、ある語に対して、その語が載っていることを予想 (期待) して辞書を引く。その予想が何度も外れると、利用者は、それ以降、その辞書を使わなくなる。つまり、辞書の見出し語集合は、利用者が、ある語がその見出し語集合に含まれるか否かを、ある程度の精度で予想できるような集合となっていなければならない。この性質を、メンバーシップ予測性と名付けよう。利用者にとって重要なことは、掲載されている期待した語が実際に辞書に掲載されていること、すなわち、見出し語集合がメンバーシップ予測性を持つことである。

我々は、辞書が満たすべき性質として、これまで、網羅性・一貫性・信頼性の 3 つを考えてきた。たとえば、『新明解国語辞典』や『岩波国語辞典』などの小型の国語辞典は、日常的に使われる語彙を収録するという点で 一貫 しており、そのほとんどを 網羅 している。同時に、その記述には誤りが少なく 信頼 できる。これら 3 つの性質のうち、網羅性と一貫性は、見出し語集合と強く関係する。本論文では、これら 2 つの性質を統括する概念として、代表性という用語を新たに導入する。その心は、「見出し語集合は、なんらかの集合群を代表する集合となっていなければならない」ということである。

小型の国語辞典においては、「日常的に使われる語彙」という集合が想定される。この集合は、想定されるだけであって、厳密に定義される (あるいは、できる) わけではない。それでもなお、このような集合を想定した議論は可能である。さて、ある人が「日常的に使われる語彙」として想定する集合 V_i は、別の人が想定する集合 V_j とは完全には一致しない。しかし、要素の大部分は重複すると考えるのが自然である。その結果、次のような「日常的に使われる語彙」群 $\mathcal{V}$ が構成されることになる。

$$\mathcal{V} = \{V_1, V_2, \cdots\} \tag{1}$$

小型の国語辞典の見出し語集合に求められることは、この集合群 $\mathcal{V}$ を代表する集合 V_* となっていることである。ここでの「代表」とは、それぞれの i に対して、積集合 $V_* \cap V_i$ に含まれる要素数が十分大きいことを意味する。以降、この集合群 $\mathcal{V}$ を見出し語集合の潜在構造と呼ぶ。

潜在構造 $\mathcal{V}$ が見出し語集合 V_* の背後に存在すると想定できる場合、V_* はメンバーシップ予測性を持つ。すなわち、利用者 i が辞書に掲載されていると期待する語 $w \in V_i$ は、高い確率で見出し語集合 V_* に含まれる。

再度確認しておくが、潜在構造 $\mathcal{V}$ は、想定されるものであり、明示的に定義されるものではない。しかしながら、見出し語集合 V_* がメンバーシップ予測性を持つということは、その背後にこのような潜在構造 $\mathcal{V}$ が想定できるということである。このことを別の言葉で言い直せば、「辞書の編纂者と多くの利用者の間で、見出し語集合に関する共通理解が成立する (あるいは、成立する前提が存在する) 場合にのみ、辞書は機能する」ということである。

4 見出し語集合の設計

では、辞書として機能するための代表性 (あるいは、メンバーシップ予測性) を有した見出し語集合は、どのようにしたら作り出せるであろうか。

まず、辞書編纂の各段階で、見出し語集合がどのように規定されるか、整理しよう。

1. 見出し語集合 V_* は、辞書が完成した暁には、要素が完全に列挙され、外延的に定義されたことになる。
2. 一方、辞書の設計段階では、V_* は、非公式に内包的に定義される（例：「日常的に使われる語彙」）に過ぎない。

このように、辞書の設計段階では、見出し語集合 V_* は厳密には定義されないが、その時点においても、背後に潜在構造 $\mathcal{V}$ を想定できるかどうかは、常識的に判断できる。たとえば、「日常的に使われる語彙」は、多くの人々(あるいは、ターゲットとする辞書利用者) が想定可能な集合であり、それらの集合は大きな重なりを持つことが期待できる。つまり、その背後に潜在構造 $\mathcal{V}$ を想定できる。その一方で、「名古屋大学で時々に使われる語彙」は、そのような潜在構造を想定できるとは考えにくい。

ここで、もう一つ、重要なパラメータが存在する。それは、見出し語集合 V_* の大きさ (要素数) である。作成しようとする辞書の見出し語集合の内包的定義を定めると、その集合の自然な (適切な) 大きさが定まる。たとえば、小型国語辞典の収録語数は 7 万語から 8 万語であり、これが、「日常的に使われる語彙」の自然な大きさである。この大きさを 3 万語にしたり、15 万語にすると、それは、見出し語の内包的定義と齟齬をきたすことになる。すなわち、見出し語集合の大きさを決めることは、見出し語集合の内包的定義 (どん

な語を収録するか) に、決定的に作用する。このため、辞書の最終設計段階では、見出し語集合の内包的定義と大きさをセットとして定める必要がある。

5 代表性の源泉

背後に潜在構造 $\mathcal{V}$ を想定できるような、適切な見出し語集合 V_* の内包的定義と大きさが定まったとしても、V_* の要素の具体的な列挙、すなわち、見出し語の選定が容易になるわけではない。例外的な場合[1]を除けば、潜在構造 $\mathcal{V}$ は人間の頭の中だけにあるとともに、V_* にはかなりの自由度がある。

一般に、代表性を有する見出し語集合を作るためには、素材情報 (見出し語候補) をどこから集めてくるか、すなわち、利用する情報源の選択が重要となる。たとえば、「日常的に使われる語彙」を見出し語集合とするのであれば、素材情報を収集する情報源として、「日常的に使われる語彙を十分にカバーする」コーパスを用いる必要がある。『現代日本語書き言葉均衡コーパス』のような均衡コーパスは、この条件を満たす。他方、適当に日本語テキストを集めてきても、その条件は満たされない。コンピュータによる自動編集では、見出し語の取捨選択に人間の常識的判断を利用できないため、代表性を持つ見出し語集合の源泉を、使用する情報源の代表性に求めざるを得ない。方法論的には、この方法が、見出し語集合に代表性を付与する唯一の方法と思われる。

他方、どんな方法で見出し語集合 V_* を作ったとして、その結果の V_* に、多くの人がメンバーシップ予測性を認めれば、その見出し語集合 V_* は代表性を持つことになる (結果としての代表性)。なお、同種のカテゴリの辞書の中で最も収録数の多い辞書は、メンバーシップ予測性が低くても、例外的に「最後の砦」の辞書として機能する[2]

6 ケーススタディ

以上のような考察との関連を中心に、現在編纂中の外国人名対訳辞書の見出し語集合をどのように決定したかについて述べる。この辞書は、人間の翻訳者が利用することを想定して設計した。辞書の編纂には、自動生成した約 35 万件のフルネーム対訳データを素材情報として使用する。この素材情報の自動生成 (収集) は、「人間の翻訳者の利用を想定した、外国人名対訳辞書を作る」という初期設計案の下で実施した。

見出し語集合の設計　最終設計に向けて採用した、見出し語集合の非公式な内包的定義は、「外国人名 (フルネーム) を構成する要素 (姓や名) の原綴のうち、主要なもの」である。Smith や John などが、その代表例となる。翻訳者の使用を想定しているので、和訳対象となる英語文書にそれなりの頻度で出現する外国人名の姓や名を、カバーしていることが望ましい。翻訳者も、同様の期待を辞書に求めると考えられるので、見出し語集合の背後に潜在構造 $\mathcal{V}$ を想定することができる。

まず、収集した 35 万件の素材情報に含まれる人名要素の異なりを調べたところ、約 10 万件という結果を得た。この素材情報は自動収集したものであるため、信頼できない人名要素も含まれる。これらを勘案し、今回作成する辞書の見出し語集合の大きさの上限を、5 万から 6 万語と定めた。[3]

この数を念頭に置き、以下に示す米国の 2 種類のデータを、見出し語数設計のための基礎資料として利用した。

1. Census (http://www.census.gov/)
   米国の国勢調査。1990 年の調査に対しては、男性名、女性名、姓の 3 種類のデータ、2000 年の調査に対しては姓のデータ、が公開されている。
2. Popular Baby Names
   (http://www.ssa.gov/oact/babynames/)
   Social Security の申請に基づく、男性名、女性名のデータ (1880 年から 2010 年)。

これらのデータに基づき、全体 (人口) の 80%をカバーするために必要な異なり数を見積もった結果、男性名 419 種類、女性名 994 種類、姓 36,767 種類となった。これは米国だけに対する見積りであり、本辞書の対象とする範囲 (ラテンアルファベットで記述される人名) は、それより広い。以上のことを考慮し、最終的に、見出し語数として、約 50,000 語という設計値を設定した。

編纂する辞書と同種の辞書に、見出し語数 91,765 件の『アルファベットから引く外国人名よみ方字典』(日外アソシエーツ, 2003) があるが、この辞書には、実例 (フルネーム) が全く記載されていない。今回編纂する辞書では、すべての姓・名に、実例を示す方針のため、収録件数が半分でも新たな価値を創出できると考えた。

[1] 新たに編纂したい辞書と同種の辞書がすでに存在する場合は、それら既存の辞書の見出し語集合を集合 V_i として利用する方法がある。辞書編纂において見出し語の剽窃が行なわれるのは、この意味で当然である。

[2] 影浦・阿辺川 [1] の comprehensiveness(包括性) という概念が、この現象と説明すると思われる。

[3] 情報パッケージの最終設計には、集めた素材情報からのフィードバック (図 1 の線 FB1) が不可欠である。

表 1: 基礎資料に対する見出し語集合の被覆率

人口比	50%		60%		70%		80%		90%	
男性名	1.00	(72/72)	1.00	(122/122)	1.00	(211/211)	1.00	(418/419)	0.92	(1537/1673)
女性名	1.00	(169/169)	1.00	(264/265)	0.98	(444/452)	0.92	(915/994)	0.58	(3068/5325)
姓	0.99	(2175/2195)	0.95	(4772/5017)	0.84	(10265/12195)	0.60	(22074/36767)	0.27	(43423/163463)

表 2: 未収録の人名要素

男性名 **(80%圏内)**	Brayden
女性名 **(70%圏内)**	Autumn, Breanna, Kaitlyn, Kaylee, Krystal, Makayla, Misty, Opal
姓 **(50%圏内)**	Bowling, Delacruz, Delatorre, Fish, Lemus, Lockett, Lyles, Madrigal, Magana, Numbers, Oakes, Painter, Peoples, Pollard, Rodriquez, Valadez, Vang, Wu, Xiong, Yee

見出し語の選定　見出し語は、上記の基礎資料と素材情報 (対訳データ約 35 万件) から、プログラムによって自動的に選定した。具体的には、次のような手順を踏んだ。

1. **暫定人名要素の設定**：基礎資料に基づき、暫定的な人名要素として 22.1 万件を認定した。
2. **信頼できる実例の選択**：素材情報から、姓名の両方が暫定人名要素である実例、22.3 万件を抜き出した。
3. **第一次見出し語選定**：抜き出した実例に含まれる人名要素の得点付けを行ない、48,648 件の見出し語を定めた。これらはすべて、暫定人名要素に含まれる。
4. **新たな実例集合の設定**：素材情報から、姓名の少なくともどちらかが上記の見出し語となっている実例、29.5 万件を抜き出した。
5. **第二次見出し語選定**：抜き出した実例に含まれる人名要素のうち、暫定人名要素に含まれないものに対し得点付けを行ない、新たに 3,862 件の見出し語を追加した。

この結果、見出し語数は 52,510 件となった。

表 1 に、自動作成した見出し語集合中に、基礎資料に基づく人口被覆率を達成するのに必要な見出し語が、どれだけ含まれているかを示す。見出し語集合は、米国の人口の 80%を被覆することを想定して設計したが、姓では、それに必要な 36,767 種類のうち、22,074 種類 (60%) しかカバーしていないことがわかる。その一方で、人口 60%を被覆するために必要な人名要素は、95%以上カバーしており、全体としては、「主要な人名要素」という内包的定義を満たす集合となっていると考えてよい。

表 2 に、選定した見出し語集合に含まれていない人名要素の例を示す。これらは、見出し語選定 (パッケージング) の時点において判明した欠落である。その原因は、素材情報収集の不備にあり、この不備は、図 1 に示したフィードバック FB2 によって補うべきである。この部分の自動化はなかなか難しいが、このような「見直し」のプロセスは、高品質の辞書の作成には、欠かすことができない。

7　まとめ

本論文の主張は、以下の 3 点に集約される。

1. 機能する辞書を作成するためには、見出し語集合の設計が不可欠である。
2. 見出し語集合は、背後に、潜在構造が想定できるものとなっている必要があり、かつ、その潜在構造の代表集合となっている必要がある。
3. 見出し語集合の設計では、内包的定義とその大きさをセットで定める必要がある。

謝辞　本研究は、科学研究費補助金基盤研究 (B)「辞書自動編纂のためのテクノロジー」課題番号 21300094 の支援を受けている。

参考文献

[1] Kyo Kageura and Takeshi Abekawa. On the concept of “comprehensiveness” in information services: The case of the online translation aid and hosting service Minna no Hon’yaku. In *Asia-Pacific Conference on Library and Information Education and Practice*, 2011.

[2] Satoshi Sato. Crawling English-Japanese person-name transliterations from the Web. In *Proc. of WWW-2009*, pp. 1151–1152, 2009.

[3] Satoshi Sato. Web-based transliteration of person names. In *Proc. of WI-2009*, pp. 272–278, 2009.

[4] Satoshi Sato. Non-productive machine transliteration. In *Proc. of Riao-2010*, pp. 16–19, 2010.

[5] Satoshi Sato and Sayoko Kaide. A person-name filter for automatic compilation of bilingual person-name lexicons. In *Proc. of LREC-2010*, 2010.

[6] 佐藤理史, 佐藤円. 情報の自動編集と WIT プロジェクト. 日本図書館情報学会研究委員会（編）, 電子図書館—デジタル情報の流通と図書館の未来, pp. 131–149. 勉誠出版, 2001.